Efficient tracking of human poses using a manifold hierarchy

Moutzouris, A., Martinez-del-Rincon, J., Nebel, J-C., & Makris, D. (2015). Efficient tracking of human poses using a manifold hierarchy. *Computer Vision and Image Understanding*, *132*, 75-86. https://doi.org/10.1016/j.cviu.2014.10.005

Published in:
Computer Vision and Image Understanding

Document Version:
Early version, also known as pre-print

Efficient Tracking of Human Poses Using a Manifold Hierarchy

Alexandros Moutzouris[a], Jesus Martinez-del-Rincon[b], Jean-Christophe Nebel[a], Dimitrios Makris[a]

[a] *Digital Imaging Research Centre, Kingston University, United Kingdom*
[b] *Queen's University, Belfast, United Kingdom*

1. Introduction

In computer vision, human motion analysis is a rapidly growing area, because of not only the methodological challenge that motion analysis implies but also its many applications such as surveillance, human computer interaction, sports analysis and computer graphics. Human motion analysis approaches have been relying on activity recognition, pose estimation and pose tracking methodologies. Whereas activity recognition aims at classifying the type of the activity performed by a human being, pose estimation deals with estimating the skeletal position of a person for one or more frames. Although this makes the solution independent of the previous poses, this is also more sensitive to errors. Consequently, pose estimation is often integrated within a pose tracking framework where past information is exploited to estimate the current pose in a more efficient way. Human pose tracking methods that rely on markerless approaches are generally desirable because of their non-invasive nature that widens significantly their potential application. Multi-camera systems are able to mitigate the complexity of markerless approaches and to deal with the inevitable limb occlusions. Still, the variety of human postures and activity styles, and the high complexity of modelling the human body make this problem a technically demanding and computationally expensive task.

Email addresses: moutzouris.a@gmail.com (Alexandros Moutzouris), j.martinez-del-rincon@qub.ac.uk (Jesus Martinez-del-Rincon), J.Nebel@kingston.ac.uk (Jean-Christophe Nebel), D.Makris@kingston.ac.uk (Dimitrios Makris)

In this paper, we introduce a hierarchical dimensionality reduction method, namely Hierarchical Temporal Laplacian Eigenmaps (HTLE). It goes beyond the hierarchical structure of human body parts, represented as pairwise relationships as in (Yang and Lee, 2006), (Urtasun and Darrell, 2008), (Amin et al., 2013) by considering divisions at different hierarchical levels, similarly to (Han et al., 2010),(Darby et al., 2009), (Raskin and Rudzsky, 2009), (Wang et al., 2011), (Tian et al., 2012). The HTLE approach allows searching each level of a posture hierarchy separately, thus modelling new, unseen poses. Furthermore, we propose a markerless, hierarchical pose tracking method, namely Hierarchical Manifold Searching (HMS), designed for multi-camera scenarios. Our framework operates in a two-phase approach; first, a training set is used in order to generate a hierarchy of low dimensional manifolds using HTLE and second, pose tracking is performed in a hierarchical manner using HMS. Therefore, unlike conventional dimensionality reduction methods which are restricted to the set of poses present in a training set, our framework is capable of moving beyond the training set and generating new poses that have never been seen before. In addition, instead of searching the whole hierarchy as performed in previous studies using computationally expensive particle filtering (Darby et al., 2009), (Raskin and Rudzsky, 2009), the computational cost of the proposed method is reduced by using deterministic optimisation applied to a subset of manifolds in the hierarchy. Our approach also deals with style variability, i.e. pose differences for a given activity resulting from either individual's personality or distinct conditions, by allowing an extra final level of hierarchy where each body part is individually adjusted in an unconstrained manner (Moutzouris et al., 2011). We show that our methodology improves computational efficiency and accuracy.

1.1. Related Work

First, different 3D human pose tracking techniques are discussed. Then, methodologies based on dimensionality reduction methods are presented. Finally, hierarchical approaches are described.

Early human pose tracking methods were based on gradient descent (Howe et al., 1999). However, they suffer from finding local optima thus giving poor tracking results. In order to search the complex space of human postures, the particle filter (PF) method (Arulampalam et al., 2002) has been used. However, the high dimensionality of this space makes it difficult to sample the solution space efficiently (Sigal et al., 2010) and prevents divergence.

Deutscher et al. (Deutscher and Reid, 2005) proposed an annealed particle filter (APF) that improves the efficiency of the particle filter search; however it is still computational expensive due to the high dimensionality. Moreover, in an unconstrained searching environment, when a particle filter-based tracker diverges, convergence in the following frames becomes problematic (Raskin et al., 2011). Gall et al. (Gall et al., 2008) introduce a multi-layer framework based on simulated annealing that combines stochastic optimization, filtering, and local optimization which led to results slightly more accurate than APF. However, as all particle based estimation methods, a large number of particles is required which increases complexity and computational cost (Sigal et al., 2010; Bandouch et al., 2008).

In order to deal with the high complexity of modelling articulated human motion, nonlinear dimensionality reduction methods have been used in tracking pipelines, exploiting available training sequences for known actions. They are grouped into two categories: mapping-based and embedded-based approaches. Mapping approaches, such as Gaussian Process Latent Variable Model (GP-LVM) (Lawrence, 2004; Hou et al., 2007), employ probabilistic nonlinear functions in order to map the embedded space to the data space. Consequently, their training is time-consuming and convergence is not guaranteed (Urtasun et al., 2007), especially for applications which are based on large training sets. Embedded approaches provide an estimate of the structure of the underlying manifold by means of approximating each data point according to their local neighbours on the manifold. The main drawback of these methods is the lack of mapping functions between high and low dimensional spaces, although Radial Basis Function Networks are usually used to resolve this issue (Lewandowski et al., 2010a). This category of techniques includes Local, Linear Embedding (Roweis and Saul, 2000) (LLE), Isometric Feature Mapping (Isomap) (Tenenbaum et al., 2000), Laplacian Eigenmaps (LE) (Belkin and Niyogi, 2001) and Local tangent space alignment (LTSA) (Zhang and Zha, 2004).

Since human motion may be described by time series, the temporal dependencies between consecutive poses can be used to improve human tracking applications. These temporal constrains ensure that points that are close in time will be close in the low dimensional space. Spatio-temporal Isomap (ST-Isomap), (Jenkins and Matarić, 2004) an extension of Isomap, changes the original weights in the graph of local neighbours in order to emphasize the similarity between temporally related points. Gaussian Process Dynamical Models (GPDM) (Wang et al., 2006) integrates time information using Gaus-

sian Process priors to create dynamics in the low dimensional space. Urtasun et al. (Urtasun et al., 2006) use GPDM for learning human poses and motion priors for 3D people tracking. However, most of these methods suffer from the fact they are person dependent: they are not able to efficient track people with their corresponding style who do not belong to the training set, which reduces their application. Alternatively, Temporal Laplacian Eigenmaps (TLE) (Lewandowski et al., 2010b) was specifically designed to address the issue of modelling activities of different people by suppressing their stylistic differences and producing a coherent manifold. The resulting manifold has a 1D dimensionality, which is suitable for fast exploration. Nonetheless, none of the above approaches allows the recovery of unseen poses. This is because dimensionality reduction methods are activity dependent, that is, they can only represent those activities that they have learned during training, usually a single activity.

Hierarchical methodologies that consider divisions of human body parts at different hierarchical levels have been proposed to extend the pose space by decoupling the motion of individual limbs which allows dealing with unseen activities. Such methodologies were proposed for 2D pose estimation in (Wang et al., 2011) (Tian et al., 2012). The Hierarchical Gaussian Process Latent Variable Model (H-GPLVM) (Lawrence and Moore, 2007) has been applied to activity recognition (Han et al., 2010) and pose estimation (Raskin and Rudzsky, 2009; Darby et al., 2009), based on a hierarchy of manifolds trained using different activities. Han et al. (Han et al., 2010) and Darby et al. (Darby et al., 2009) used H-GPLVM for training two different activities and the APF method to search for poses that result from combinations of these activities. Using this learnt hierarchical model for multiple activities they can recover novel poses which are not present in the training dataset. For example, training data for a person walking and a person standing and waving allow detecting a person who is walking whilst waving. The hierarchy is able to recognise the posture of the upper body from the first training activity and that of the lower body from the other one. Similarly, Raskin et al. (Raskin and Rudzsky, 2009) presented an extension of the Gaussian Process Annealed Particle Filter (GPAPF) method (Raskin et al., 2011) called Hierarchical Annealing Particle Filter (H-APF). This method also uses H-GPLVM to generate a hierarchy of manifolds in the low-dimensional space, and the APF method to generate particles in the latent space. H-APF is tested in a multi-activity scenario combining walking and jogging activities, where the activity of every frame is estimated before pose estimation. Al-

though all these hierarchical approaches allow the generation of unseen poses where individual body part postures originally belonged to different activities, their main drawback is their high computational cost, since APF is used to search through the whole hierarchy.

1.2. Overview

The pipeline of our approach is presented in Figure 1. More specifically, the training set comes from MoCap data describing the activity of interest as a sequence of human poses. Activity Manifolds are learned by the proposed Hierarchical Temporal Laplacian Eigenmaps (HTLE) (Figure 1a), as described in section 2. The pose tracking process is constrained by the hierarchy of activity manifolds (Figure 1b) which is presented in section 3. Our system (Figure1b) assumes multiple calibrated and synchronised cameras . At every cycle, an observation is estimated from the set of images captured by the vision system. The observation and the previously learnt Activity Manifolds are fed to our novel search method, i.e. Hierarchical Manifold Search (HMS), which explores efficiently the pose space described by HTLE. An observation is proposed based on volume overlap between the observation and a 3D geometric human model, and on colour information of the input data. The final output is the 3D coordinates of the joints of the estimated pose for each set of synchronised frames.

2. Activity Manifold Learning

In this section, we present the formation of Hierarchical Temporal Laplacian Eigenmaps (HTLE). Since TLE generates a coherent low dimensional manifold that takes into account the temporal dependencies of the data, it can only model poses seen in the training dataset. In order to deal with this restriction we propose to expand the available pose space using HTLE, a hierarchy extension of TLE. The advantages of such structure are two-fold: firstly, fast searching is facilitated by a set of 1D TLE manifolds; secondly, the hierarchy of manifolds models unseen poses to address the problem of variations between the subjects of the training and the testing datasets. After a presentation of the TLE dimensionality reduction method, HTLE is described in details.

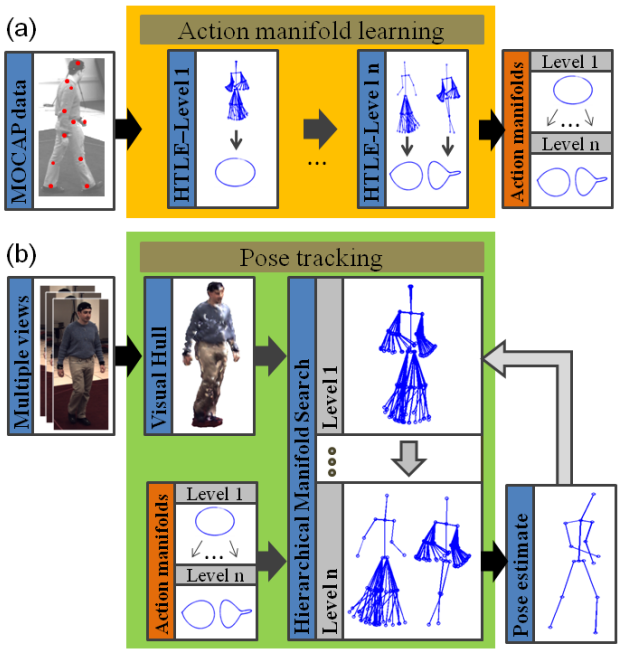

Figure 1: (a) Training and (b) pose tracking pipelines.

2.1. Temporal Laplacian Eigenmaps (TLE)

The Temporal Laplacian Eigenmaps dimensionality reduction method has been applied to represent sequences of human poses for a given activity (Lewandowski et al., 2010b, 2014). TLE generates a temporal representation of human postures, where inter-person variability (style) has been suppressed, expressed as a one-dimension manifold. Here, we have selected TLE as the dimensionality reduction method for two reasons. Firstly, TLE explicitly preserves the temporal coherence of an activity, which is important for a tracking application. Secondly, searching one-dimensional manifolds, such as those produced by TLE, is very efficient.

In order to generate the low dimensional manifold, a training dataset P is used

$$P = \left\{p^i, i = 1, ..., N\right\}, p^i \in \mathbb{R}^D \tag{1}$$

corresponds to N poses, where p^i is the ith pose of the model. TLE produces a manifold Q, which is an equivalent representation of P in a low dimensional space of d dimensions,

$$Q = \left\{q^i, i = 1, ..., N\right\}, q^i \in \mathbb{R}^d \tag{2}$$

where $D, d \in \mathbb{N}$, $d \ll D$ and q^i is the representation of pose p^ion the manifold.

The manifold Q is calculated by defining firstly a neighbourhood graph and, then, solving the eigenvalue problem. Since TLE aims to preserve temporality, two types of temporal neighbourhoods are defined for each data point p^i. Adjacent temporal neighbours A: the $2m$ closest points in the sequential order of input (Figure 2a)

$$A^i \in \left\{p^{i-m}, \ldots, p^{i-1}, p^i, p^{i+1}, \ldots, p^{i+m}\right\} \tag{3}$$

and repetition temporal neighbours R: the s points similar to p^i, extracted from repetitions of time series fragment F^i (Figure 2b)

$$R^i \in \left\{p^{i,1}\left(C\right), \ldots, p^{i,s}\left(C\right)\right\} \tag{4}$$

where $p^{i,j}\left(C\right)$ returns the centre point of an activity fragment similar to F^i for each iteration j. F^i is a fragment of the activity sequence defined by the central point $p^i$and a fixed number of surrounding adjacent temporal neighbours. Once the frament F^i is defined, the repetitions of this fragment, that is, $F^{i,1}, ..F^{i,s}$, are extracted by applying Dynamic Time Warping over the full dataset. More details about this process and its parameters can be found in (Lewandowski et al., 2010b, 2014).

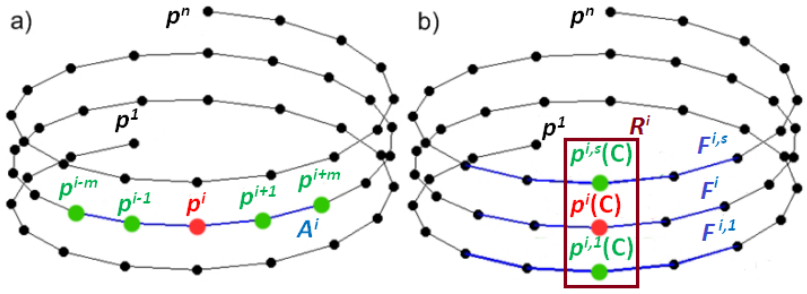

Figure 2: Adjacent temporal a) and repetition temporal b) neighbours (green dots) of a given data point, p^i, (red dots).

Once the neighbourhoods for each data point are defined, the neighbourhood graph is estimated. First, using the standard LE formulation the weights W

$$W^G_{i,j} = \begin{cases} e^{\left\|p^i - p^j\right\|^2} & i, j \text{ connected} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

are assigned to the edges of each graph $G \in \{A, R\}$.

Then, following the standard LE formulation, an extended cost function is introduced to combine information from both graphs

$$\arg\min_{Q} Q^T \cdot (L_A + L_R) \cdot Q \tag{6}$$

$$\text{subject to:} Q^T \cdot (D_A + D_R) \cdot Q = I \tag{7}$$

where $D_G = diag\left\{D_{11}^G, D_{22}^G, ..., D_{nn}^G\right\}$ is a diagonal matrix with $D_{ii}^G = \sum_{j=1}^{n} W_{ij}^G$, and $L_G = D_G - W^G$ is the Laplacian matrix corresponding to the graph. The minimum of the objective function can be found by solving the eigenvalue problem from the equations

$$\wedge(Q, \lambda) = Q^T (L_A + L_R) Q - \lambda \left(I - Q^T (D_A + D_R) Q\right) \tag{8}$$

$$(L_A + L_R) Q = \lambda (D_A + D_R) Q \tag{9}$$

The generalised eigenvalue problem is solved to span the embedded space Q by the eigenvectors given by the d smallest nonzero eigenvalues λ.

Unlike the standard Laplacian Eigenmap dimensionality reduction method (LE) that only preserves the manifold's local geometry (Belkin and Niyogi, 2003), the temporal structure of the data manifold is preserved thanks to the inclusion of the graphs A and R. Consequently, TLE is able to preserve implicitly the local and global temporal topology of the data. This implies that TLE maintains the temporal continuity of time series during dimensionality reduction process and suppresses stylistic variations displayed by different sources of time series by aligning them in the low dimensional space (Lewandowski et al., 2010b).

Although the manifold lies in the low dimensional space, the observation function needs to be evaluated in the high dimensional space. Consequently, a mapping function is required to find correspondences between the two spaces. Since spectral methods lack mapping functions to project data from one space to another, Radial Basis Function Network (RBFN) as suggested by (Lewandowski et al., 2010a) are trained to obtain these transformations φ and φ':

$$\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^d \text{and } \varphi' : \mathbb{R}^d \rightarrow \mathbb{R}^D \tag{10}$$

TLE have been used for modelling a wide variety of activities, both periodic and non-periodic (Lewandowski et al., 2010b, 2014).

2.2. Hierarchical Temporal Laplacian Eigenmaps (HTLE)

The hierarchical structure of HTLE dimensionality reduction method has been designed to allow searching each level of the hierarchy extending the training dataset. This is achieved by exploring each level separately and then combining all of them, generating a new, unseen configuration. In this study, we use HTLE for pose tracking purposes as the human body can be regarded as a hierarchical structure.

HTLE uses the training dataset P to generate a hierarchy of manifolds in low dimensional spaces. Let $P_{h,l}$ be the set of N poses of the training dataset that corresponds to the l-th pose subspace at the hierarchy level h

$$P_{h,l} = \left\{ p_{h,l}^{i}, i = 1, ..., N \right\}, \tag{11}$$

where $p_{h,l}^{i} \in \mathbb{R}^{D_{h,l}}$ is the pose of the model at the time i. As discussed in Section 2.1 TLE produces a manifold $Q_{h,l}$ representing $P_{h,l}$ in a low dimensional space $\mathbb{R}^{d_{h,l}}$

$$Q_{h,l} = \left\{ q_{h,l}^{i}, i = 1, ..., N \right\}, \tag{12}$$

where $q_{h,l}^{i} \in \mathbb{R}^{d_{h,l}}$ and $d_{h,l} \ll D_{h,l}$.

At a given level h (Figure 3), mapping between the high and low dimensional spaces (Lewandowski et al., 2010b) is performed by the functions:

$$\varphi_{h,l} : \mathbb{R}^{D_{h,l}} \rightarrow \mathbb{R}^{d_{h,l}}, \varphi'_{h,l} : \mathbb{R}^{d_{h,l}} \rightarrow \mathbb{R}^{D_{h,l}} \tag{13}$$

where

$$\varphi_{h,l}\left(p_{h,l}^{i}\right) = q_{h,l}^{i}, \varphi'_{h,l}\left(q_{h,l}^{i}\right) = p_{h,l}^{i}. \tag{14}$$

We also define mapping functions (Figure 3) between the hierarchy levels points $p_{h-1,l} \in P_{h-1,l}, p_{h,l'} \in P_{h,l'}$

$$\omega_{h,l'} : P_{h-1,l} \rightarrow P_{h,l'}, \text{where } \omega_{h,l'}\left(p_{h-1,l}\right) = p_{h,l'} \tag{15}$$

These mapping functions permit evaluating hypotheses by projection to the high dimensional space as well as propagating hypotheses through the hierarchy.

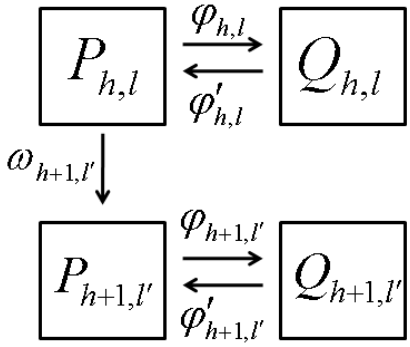

Figure 3: Pose subspaces P and submanifolds Q connected by mapping functions $\varphi, \varphi'$and ω.

3. Pose Tracking Framework

In this section we introduce the Hierarchical Manifold Search (HMS) method. HMS explores a hierarchy representing the human body as created by HTLE to estimate a human pose. For each frame, at the first level of the hierarchy, the optimal pose is estimated by searching in a low dimensional manifold representing the full body motion of the activity of interest. Then, the subspaces associated to body parts are searched hierarchically looking for the optimal pose at a given level. To achieve this, we create the observation from the input images of the dataset. Then, the observation function compares in an intuitive manner the observation H, and pose hypotheses generated by the human model M. Finally, the estimated human pose is fed back as input for the next frame.

3.1. Pre-processing

From the input data an observation is generated for every frame. Since our input may be acquired by multiple synchronised sensors, the term frame may also mean a set of synchronised frames in this work. The observation includes the information that will be used in the pose tracking methodologies.

A $3D$ volumetric representation (observation) of the observed human is generated to allow evaluation of human model hypotheses. Specifically, we assume that the testing set comprises synchronised views of a human from multiple cameras. A background subtraction method is needed to extract the human body silhouette from every view; i.e. the constant background is removed from every image and the result is the body silhouette including

the colour information. In this work, the standard background subtraction method suggested by HumanEva (Stauffer and Grimson, 1999) is used to ensure fair comparison with other methods.

Then, the silhouette images are used to generate the observation. When the foreground is projected into the $3D$ space, using the calibration information of the camera, a $3D$ geometric shape is defined that contains the target object. The observation is generated from the intersection of all silhouette geometric shapes, and represented as a set of $3D$ voxels. In this study the Bounding Edge technique (Cheung et al., 2005) is used to generate the Visual Hull. This procedure has been refined since our previous work (Moutzouris et al., 2012): the colour from the input images is also back-projected on the visual hull (Fitzgibbon et al., 1998) (Figure 4) in order to discriminate between body parts and improve accuracy.

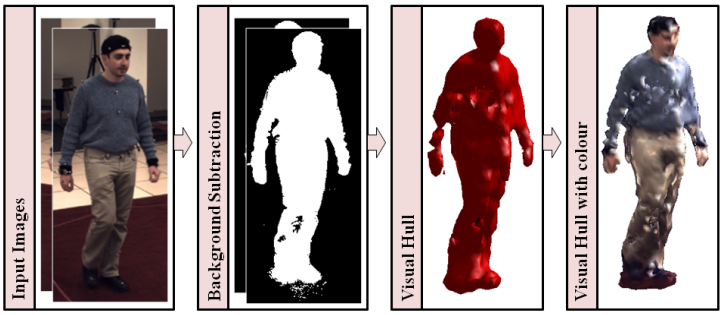

Figure 4: The pre-processing pipeline. From left to right: the input images, the corresponding silhouettes, the visual hull and the visual hull with colour.

3.2. Human Model

For the purpose of our method we use a $3D$ articulated human model M that consists of L cylindrical parts. The cylindrical definition of the model allows us to compare pose estimates against the observation using simple mathematical models. This is for faster evaluation of the observation function without losing the basic structure of the 3D human shape.

The human model M is defined as a set of three independent parameters

$$M = \left\{ \{g, p, m\}, g \in \mathbb{R}^6, p \in \mathbb{R}^D, m \in \mathbb{R}^{2L} \right\} \tag{16}$$

where $g \in \mathbb{R}^6$ describes the global rotation and translation of the body into the $3D$ Euclidean space, $p \in \mathbb{R}^D$ the pose of the model that is expressed by

joint angles between body parts and $m \in \mathbb{R}^{2L}$ represents the human volumetric model expressed by the length and the radius of the cylinders of the L body parts. Joint angles are represented by quaternions, as a consequence of which every body part requires four parameters , i.e. $D = 4 \cdot L$.

In tracking experiments, the human model M is initialised manually in the first frame. While m is considered fixed for every human subject, tracking involves recovering the global position/orientation g and the joint angles p at every frame. The Skeleton Representation is a model that is extracted from specific points of the Volumetric Representation. Every part of the human body is represented by a straight line that connects two points as seen in Figure 5.

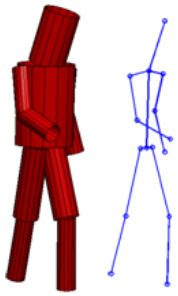

Figure 5: Human model and corresponding skeleton representation.

3.3. Observation Function

In order to compare the observation with the pose hypotheses, an observation function is used,It based on the coloured visual hull H (Cheung and Others, 2003) and consists of two parts. The first part s_1 compares the $3D$ volume of the observation H with the pose hypothesis M by using the relative overlap between them:

$$s_1\left(M, H\right) = \frac{\mid M \cap H \mid}{\mid M \mid}. \tag{17}$$

The second part s_2 is based on the colour similarity. This is important since it complements the first part of the observation function especially for poses where the limbs are close to the torso as the colour of the torso is often different from the limbs'. At the initial pose the colour of the limbs $c_j^1, j = 1, ..., L$ is estimated, using the voxels of the initial observation H^1,

matched by the limb j. Then, this colour information is used for comparing corresponding areas of this frame with the initial one. Specifically, $c_j^1, j = 1, ..., L$ is estimated as the average of the hue values of all the matched voxels, assuming an HSV colour space. The hue value is used for comparing the colour without affecting the saturation and the brightness at every frame due to shadows and light. Then, at the frame i the colour information of the observation H^i of every voxel v, $c_j^{i,v}, j = 1, ..., L$, matched by the limb j, is compared to the initial limb colour c_j^1. A binary colour similarity variable, $C_j^{i,v}$, is introduced to emphasise significant differences and at the same time suppress noise in the hue channel

$$C_j^{i,v} = \begin{cases} 1, & \text{if } \left|c_j^{i,v} - c_j^1\right| \leq a \\ 0, & \text{if } \left|c_j^{i,v} - c_j^1\right| > a \end{cases} \tag{18}$$

where $a \in [0, 1]$ is an appropriate threshold. Then the observation function s_2 is defined by:

$$s_2\left(M, H\right) = \frac{1}{L} \sum_{j=1}^{L} \frac{\sum_{v=1}^{V_j} C_j^{i,v}}{V_j} \tag{19}$$

where V_j is the total size in voxels of limb j, $C_j^{i,v}$ is a binary variable, that emphasises significant colour similarities and L is number of the body parts.

The observation function f is given by the weighted mean

$$f\left(M, H\right) = \sum w_n s_n(M, H) \tag{20}$$

where w_n is the weight that allows changing the balance between observation functions, where $\sum w_n = 1$.

An advantage of the proposed observation function is that it allows comparisons of individual body parts of the human model to the observation as seen in Figure 6. This property is important when moving down through our hierarchy in 2.2. Also, because of the $3D$ representation, individual body parts, like torso or arms, may be removed from the observation without affecting the observation of other body parts, making the search strategy more efficient, as explained later in section 3.5. This contrasts with 2D image-based observation functions, such as the silhouette and edge likelihood and the bi-directional silhouette likelihood that are tested in (Sigal et al., 2010) that do now allow comparison of individual body parts because of potential occlusions in image views.

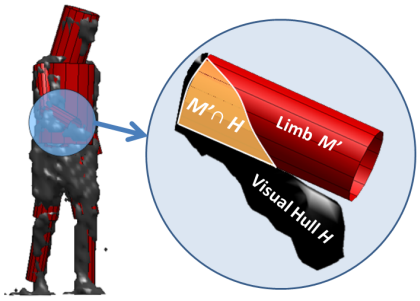

Figure 6: Calculation of observation function s_1 for individual body parts.

3.4. Learning

HTLE is selected to model the pose search space for two reasons. Firstly, the one-dimensionality of each manifold facilitates fast searching. Secondly, its hierarchical structure provides flexibility to the search space to allow detection of unseen poses. We define a hierarchy based on the division of the individual body parts as shown in Figure 7. At the first level h_1, the whole body is represented. At the next level h_2, the variability of the previous level is expressed by two subspaces containing either the upper or the lower body. The division process is repeated for the next two levels h_3 and h_4: firstly, four subspaces are created to model the four individual limbs, i.e. left and right arms and legs; secondly, each limb is divided into two segments, i.e. upper and lower arm and leg, to produce eight submanifolds. At the last level h_5, each limb segment is allowed to move in an unconstrained manner similarly to (Moutzouris et al., 2011). Although levels h_4 and h_5 correspond to the same leaf nodes, their search spaces are different since only h_4 is constrained by the training dataset. Nonetheless, we include both in the hierarchy for simpler representation of the pose tracking method. By introducing different levels with an increasing level of specificity, we incrementally vary the ability of generating new pose hypotheses while maintaining a certain level of constraints.

3.5. Pose Tracking

In this section, we introduce the Hierarchical Manifold Search (HMS) method, which is used to estimate the human pose through the hierarchy proposed in 2. Initially, we search the top level of the hierarchy, which represents the full body pose. Then, if the result of the observation function

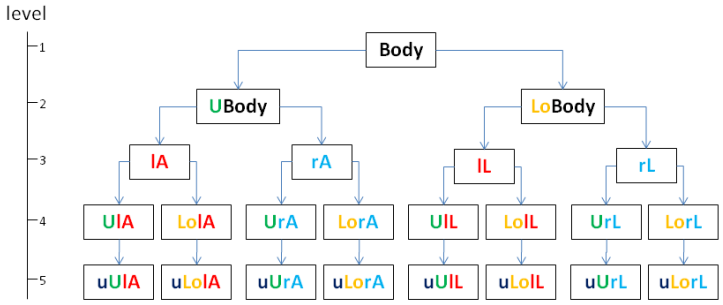

Figure 7: Five-level hierarchy of human model. Each level is represented horizontally in the figure. Level number increases by one progressively from top to bottom. Every level h is composed of pose subspaces l. U: Upper, Lo: Lower, l: left, r: right, A: Arm, L: Leg, u: unconstrained

is not satisfactory, we search incrementally the remaining levels of the hierarchy, each of them representing a different division of the human body. This procedure allows taking full advantage of the hierarchy of manifolds which mitigates discrepancies between the testing and training datasets by permitting the estimation of unseen poses.

For every frame i, we optimise the observation function $f\left(\{g^i, p^i, m\}, H^i\right)$ in two steps. Firstly, we initialise the global position and orientation g^i of the human model with the previous frame p^{i-1}. The observation H^i generated for every frame i is compared with a human model hypothesis $M^i = \{g^i, p^{i-1}, m\}$ by maximising the observation function $f\left(M^i, H^i\right)$, varying the global parameters g^i

$$\dot{g}^i = \left\{g^i : \max f\left(M^i, H^i\right)\right\}. \tag{21}$$

The voxels that are spatially matched to the torso of M^i are removed from the observation H^i. This allows faster evaluation of the observation function and also setting body constraints such as the arm not going through the body to avoid errors in the estimation of the limbs that are near the torso. Since the goal is to evaluate the hypotheses M^i, failure in torso estimation is penalised giving a lower score compared to a hypothesis with a well-located torso. If the position of the torso is accurate then by removing the torso voxels, errors are avoided in the estimation of the limbs that are near the torso. If the position of the torso is not accurate, then it also introduces a similar error in the cost function even if the torso is not removed.

Secondly, the pose p^i of the current frame i is estimated. Specifically, a process is applied through the hierarchy, as illustrated in Figure 8. We apply

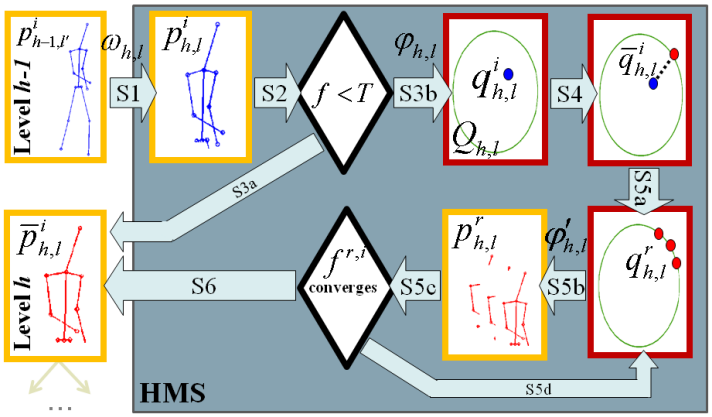

Figure 8: Flowchart of HMS at subspace (h, l) of the hierarchy. Transformations in the high and low dimensional spaces are represented in orange-framed and red-framed boxes, respectively.

the following algorithm to each TLE manifold l, for each TLE-constrained level h.

Initially, a new hypothesis $p^i_{h,l}$ for frame i is generated (Figure 8, S1). If $h = 1$, the pose from the previous frame is projected to the pose subspace $P_{1,1}$, i.e.

$$p^i_{1,1} = p^{i-1} \tag{22}$$

otherwise if $h > 1$ the point from the pose subspace l' , from the previous hierarchical level $h-1$ is projected to the child pose subspace $P_{h,l}$ using the projection function $\omega_{h,l}$ (Eq.15) to restrict the part of the human model that is searched:

$$p^i_{h,l} = \omega_{h,l}\left(p^i_{h-1,l'}\right) \tag{23}$$

Then, the model hypothesis is compared to the observation using the observation function (Figure 8, S2) (Eq.20).

If the match is sufficiently large (Figure 8, S3a), i.e.

$$f\left(p^i_{h,l}, H^i\right) > T, \tag{24}$$

where T is linked to the required accuracy, searching the current subspace (h, l) and its child subspaces in the hierarchy $(h+1, l), (h+2, l), ...$ is omitted. Therefore, the final estimation for this subspace is given as: $\dot{p}^i_{h,l} = p^i_{h,l}$ and

HMS proceeds with the manifolds of the next body parts $l+1$ and levels $h+1$ (S1).

Otherwise, the hypothesis is assumed unsatisfactory and the high dimensional point $p^i_{h,l}$ is projected to the low dimension space $\mathbb{R}^{d_{h,l}}$ to find a better estimate (Figure 8, S3b):

$$q^i_{h,l} = \varphi_{h,l}\left(p^i_{h,l}\right). \tag{25}$$

Then, the solution is constrained using the activity manifold that represents the articulate poses of the training dataset. Specifically, HMS initialises the manifold search with the closest point $\dot{q}^i_{h,l}$ on the manifold to the point $q^i_{h,l}$ in $Q_{h,l}$ (Figure 8, S4).

Afterwards, the local maximum is searched by optimising the observation function on the manifold surface. A gradient descent optimisation algorithm is used in order to find a local maximum where putative solutions are evaluated and scored in the high-dimensional space using the observation function. More specifically, this is achieved by following the four following sub-steps (Figure 8, S5).

A point $q^r_{h,l} \in Q_{h,l}$ is selected using a gradient-based optimisation algorithm (Figure 8, S5a). The point $q^r_{h,l}$ is back-projected to the high dimensional space $\mathbb{R}^{D_{h,l}}$ of human pose (Figure 8, S5b). Let $p^r_{h,l}$ be the point after the projection

$$p^r_{h,l} = {\varphi'}_{h,l}\left(q^r_{h,l}\right) \tag{26}$$

The observation function of the point $p^r_{h,l}$ is estimated (Figure 8, S5c):

$$f^{r,i}_{h,l} = f\left(p^r_{h,l}, H^i\right) \tag{27}$$

The search continues (Figure 8, S5d) until the observation function converges to a solution. Finally, the output of the algorithm is the optimal point $\dot{p}^i_{h,l}$ that maximises the observation function $f^{r,i}_{h,l}$ (Figure 8, S6)

$$\dot{p}^i_{h,l} = \left\{p^r_{h,l} : \max_r f^{r,i}_{h,l}\right\} \tag{28}$$

At the last level h' of the hierarchy, Limb Correction may be applied to refine the solution in an unconstrained space only if no satisfactory solution is found through searching all the previous levels of the hierarchy, according to the threshold T (Moutzouris et al., 2011). This process is decomposed in five steps and depicted in Figure 9.

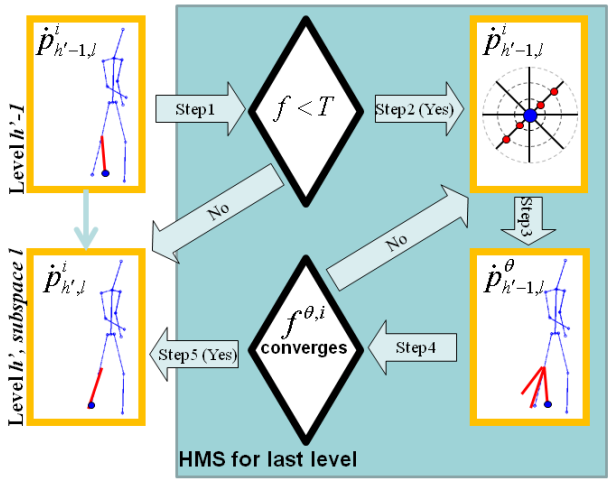

Figure 9: Flowchart of HMS for the last level of the hierarchy.

(Figure 9, Step1) The hypothesis of the limb $\dot{p}^i_{h'-1,l}$ is compared to the observation using the observation function. If the pose $\dot{p}^i_{h'-1,l}$ derived through searching in $P_{h'-1,l}$, is not satisfactory according to the threshold T , i.e.:

$$f\left(\dot{p}^i_{h'-1,l}, H^i\right) < T \tag{29}$$

then we further search for the optimal solution in the high-dimensional limb pose space and proceed to Step2. Otherwise, the current limb estimate is considered to be sufficiently accurate.

(Figure 9, Step2) Then a deterministic optimisation method is applied to detect the optimal position of the limb. We search for the rotation angle θ of the limb that optimises the observation function for the limb $\dot{p}^i_{h'-1,l}$. Since the solution space may be represented by the surface of a sphere that comprises all the possible position of the limb rotating around the articulation joint, searching is performed on that surface: the point $\dot{p}^i_{h'-1,l}$ is selected using a gradient-based optimisation algorithm.

(Figure 9, Step3) The observation function of the limb pose $\dot{p}^i_{h'-1,l}$ is estimated:

$$f^{\theta,i} = f\left(\dot{p}^\theta_{h'-1,l}, H^i\right) \tag{30}$$

(Figure 9, Step4) The estimated pose is fed back to Step 3 until the observation function converges to a solution.

(Figure 9, Step5) After maximising the function $f^{\theta,i}$ in steps 3 and 4, the final pose estimate $\dot{p}^i_{h',l}$ is the output for limb j.

$$\dot{p}^i_{h',l} = \arg\max_{\theta} f^{\theta,i} \tag{31}$$

At the end, the pose of the model p^i is estimated by concatenating the body parts estimated at the last level of the hierarchy.

$$p^i = \left\{\dot{p}^i_{h',1}, \dot{p}^i_{h',2}, ..., \dot{p}^i_{h',n}\right\} \tag{32}$$

Thus, the estimated human model M^i is

$$M^i = \left\{m, g^i, p^i\right\} \tag{33}$$

where g^i is the global position and m is a known matrix from the initial frame. Level of freedom between body parts is gradually increased (and at the same time the level of coordination decreases) through the hierarchy since "range of motion from one limb is directly influenced by the pose of the connecting structure" and decided dynamically for every pose when there is an indication that keeping the coordination of the limbs leads to unsatisfactory fitting to the observation.

HMS allows a data-driven efficient search of the hierarchy of manifolds, compared to previous hierarchical approaches, (Raskin and Rudzsky, 2009; Darby et al., 2009). The threshold T controls this search, i.e. the lower the threshold, the lower accuracy is needed, the faster the search is performed and the least the solution deviates from the training dataset, as shown in section 4. Although our approach is based on gradient-descent optimisation, the hierarchy structure minimises the problem of being trapped into a local optimum, by searching again limb configurations at different levels, as shown in the results presented later.

4. Experimental Results

In this section, HMS method is validated using publicly available datasets and compared with state-of-the-art human pose tracking methods. The sensitivity of the method to the different parameters, such as the number of hierarchy levels or accuracy thresholds is also analysed in order to calculate the trade-off between computational cost and accuracy.

4.1. Datasets

We evaluate HMS on publicly available dataset. Specifically, we use the multiple activity sequences of the HumanEva (HE) II dataset (BrownUniversity, 2007), i.e. HEII-S2 frames 1 to 710 ($1-390$ walking, $391-710$ jogging), HEII-S4, frames 4 to 710 ($4-370$ walking, $371-710$ jogging) and Image & MOCAP Synchronized Dataset (IMS) (Sigal, Leonid, 2004). For all sequences we used human activities captured by 4 cameras and calibration information for each of them. In all experiments, the tracker is initialised with the first frame of the sequence using the ground truth pose provided by the dataset. Using the ground truth pose the human model (3.2) is generated. The standard metric proposed by Sigal et al. (Sigal et al., 2010) is used to quantitatively evaluate the results. Error is calculated for each of the 15 points of the skeleton representation as the Euclidean distance between the point of the estimated skeleton and the corresponding point of the ground truth.

4.2. Training

A training dataset is used to generate the HTLE models as discussed in 2.2. Training and testing datasets are always different. For the HumanEva dataset, walking and jogging HTLE models are estimated using 1443 skeleton poses from the HEI-S2 walking, trial-3 and 795 skeleton poses from the HEI-S2 jogging, trial-3 sequences respectively. The same training dataset is used for all experiments for each activity to demonstrate the generalisation properties of the HMS method. In Figure 10 the human poses that correspond to the training data set $P_{h,l} \in \mathbb{R}^D$ and the corresponding manifolds in $2D$, $Q_{h,l} \in \mathbb{R}^2$ are shown for different levels of the hierarchy h and pose subspace l.

4.3. HMS Configuration

In this section, we investigate different configurations of the HMS method by evaluating different sets of levels in the hierarchy and different values of the threshold T.

Figure 11 shows the average error and the computational time per frame for 150 frames of the IMS dataset for different HMS configurations. As shown in Figure 11(a) by increasing the levels of the hierarchy, the estimated error decreases for every threshold. Furthermore, for increasing threshold the error decreases in all configurations. Likewise, as shown in Figure 11(b), computational cost (mean number of observations per frame for all frames) grows with

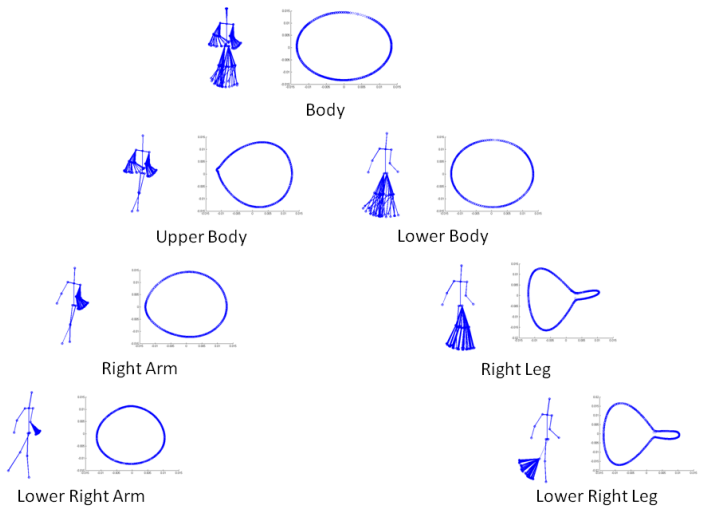

Figure 10: Different levels of the hierarchy. Human poses and the corresponding manifolds are represented in 2D.

increasing levels of hierarchy since it leads to a higher number of subspaces that are searched in every level as seen in Figure 7. Finally, computational cost increases for increasing thresholds. Figure 11(c) shows the mean number of observation evaluations per frame for different thresholds and levels of the hierarchy in the HMS(1-5) configuration. The mean number of observations per frame for every level increases for increasing threshold. Therefore, different configurations of HMS provide flexibility on compromising between computational cost and accuracy, demonstrating the value of the hierarchy.

Once the threshold reaches 0.8, error and computational cost are almost constant since the maximum value of the observation function is near 0.8. Finally, computational cost depends on many parameters. For these experiments we used an IntelCore 2 laptop with unoptimised code written in Matlab. Computational costs vary from $4sec$ to $55sec$ per frame.

4.4. Evaluation and Comparison of HMS

In order to compare the HMS method with state-of-the-art methodologies we apply HMS to the Walking activity of HEII-S2 (frames 1 to 390), HEII-S4 (frames 4 to 297) and HEI-S1walking1 (frames 1 to 590) and to the Jogging activity of HEII-S2 (frames 391 to 710) and HEII-S4 (frames 371 to 790). For every activity we use the corresponding training dataset as discussed in section 4.2. For all sequences, 4 cameras are used and the ground truth for

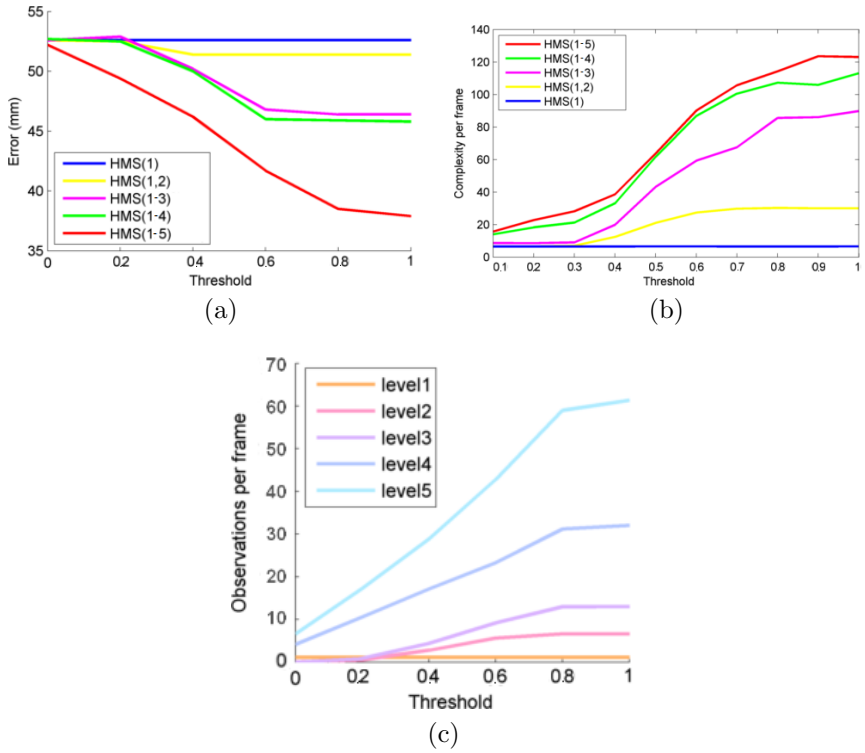

Figure 11: HMS performance for different thresholds and configurations (different numbers of hierarchy levels). (a) Average error of different configurations of HMS for 150 frames and different threshold $(0-1$ and (b) average number of evaluations of the observation function per frame for HMS method for increasing threshold $(0-1)$. (c) Mean number of observation evaluations per frame for different levels of the hierarchy and different threshold in the HMS(1-5) configuration.

the first frame initialises the tracker. Since ground truth is not known for the full length of the sequences, the results of the HMS method were evaluated using the online evaluation system of the Human Eva website (BrownUniversity, 2007). Our method is quantitatively evaluated against MP (Manifold Projection) method, MPLC (Manifold Projection Limb Correction) method (Moutzouris et al., 2011), APF that demonstrates state-of-the-art performance according to (Sigal et al., 2010; Balan et al., 2005) and applications of APF in low-dimensional spaces, i.e. GPAPF (Raskin et al., 2011), H-APF (Raskin and Rudzsky, 2009).

In Table 1 we present the average absolute 3D error, (Sigal et al., 2010), for GPAPF and MP, and the corresponding hierarchical methods, i.e. H-APF

	HEIIS2walk	HEIIS4walk	HEIS1walk	Comp.
GPAPF	86.6 mm	89.0 mm	86.3 mm	500 obs/f
H-APF	75.2 mm	81.8 mm	75.4 mm	500 obs/f
MP	74.0 mm	96.2 mm	72.0 mm	10 obs/f
MPLC	71.4 mm	75.6 mm	68.8 mm	60 obs/f
HMS	63.1 mm	62.5 mm	65.0 mm	130 obs/f

Table 1: Average error in mm and complexity (Comp.) (mean number of observations per frame) for GPAPF, H-APF, MP, MPLC and HMS methods.

and HMS. We also present the complexity (mean number of observations per frame) for every method. In this experiment, a threshold $T = 1$ (Eq.24) is set for HMS to achieve optimal results. These results demonstrate the value of introducing hierarchy in dimensionality reduction based approaches, as hierarchical methods performed better than the original ones, and improved computational efficiency and accuracy compared with GPAFP and H-APF. Our decision to base our dimensionality reduction framework on TLE is confirmed by the comparison between TLE-based and GPLVM-based representations. Specifically, MP and HMS outperforms in most of the cases GPAPF and H-APF, respectively. In Table 1 we also compare the MPLC method with HMS. HMS outperforms the MPLC method in all cases.

In Figure 12 we show the average error per frame for HE-II S2 walking and HE-II S4 walking datasets for MP (blue line) and HMS(1-5) (red line) methods using threshold $T = 1$. HMS(1-5) clearly improves over MP in all datasets (see Table 1). This confirms the value of using the hierarchy.

Figure 13 displays the average absolute 3D error for APF and HMS using different particle numbers and thresholds respectively and their computational costs as measured on the same machine using Matlab implementations for both methodologies. HMS using $T = 1$ generally outperforms APF both in terms of error and complexity. Moreover, the figure suggests that HMS is able to deliver similar accuracy to any APF configuration using only $5\% - 25\%$ of processing time. The low complexity of our method comes from the hierarchical searching strategy that is driven by the observation function. Furthermore, the combination of a hierarchical approach with a search that occurs beyond the training dataset results in improved accuracy. In summary, HMS methodology achieves the best overall accuracy with the lowest computational complexity.

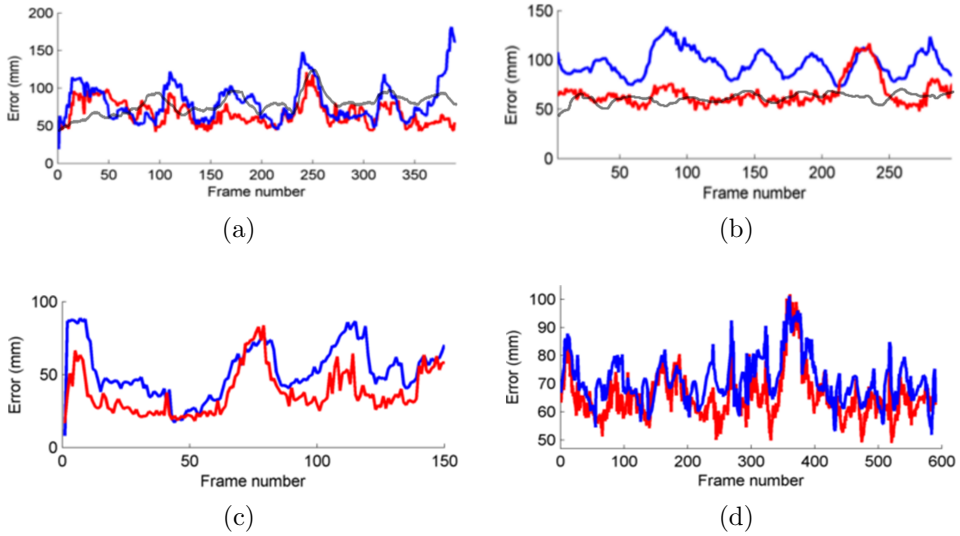

Figure 12: Results for (a) HEII-S2, (b) HEII-S4, (c) IMS and (d) HEI-S1walking1 sequence with MP (blue line), HMS(1-5) (red line) and APF (black line) methods when available.

In Figure 14 we show the average error per frame for HE-II S2 walking and jogging dataset using HMS(1-5) for lower body (red line), upper body (blue line) and full body (black line). The training dataset is the walking activity as described in 4.2. The average error for the lower body is $63.7mm$ ($57mm$ for walking and $70.4mm$ for jogging), for the upper body is $96.2mm$ ($69.2mm$ for walking and $123.2mm$ for jogging) and for the full body is $79.6mm$ ($63.1mm$ for walking and $96.7mm$ for jogging). As expected, error in the walking sequence (frames $1-390$) is lower than that of the jogging activity (frames $390-710$), since training was based on walking data. More specifically, error in the jogging activity is higher mainly because of upper body error: in the tested jogging activity, arm positions are significantly dissimilar from those found in the walking dataset, especially when arms are near to the torso. Since the latter configuration is periodical over the jogging activity, a cyclic pattern of error is observed in Figure 14. On the other hand, although a walking activity was used for training, leg positions were estimated accurately for both walking and jogging activities. These results suggest that our methodology is able to track different styles efficiently to the extend that these are not significantly dissimilar to the training set, so they can still be considered as a variation of the same given activity.

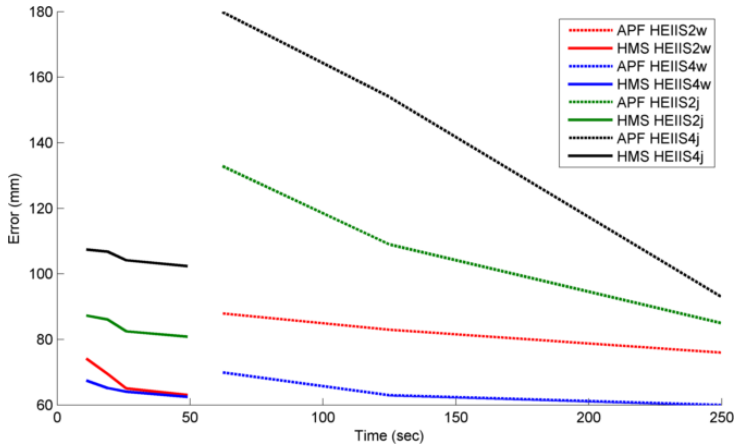

Figure 13: Average error in mm and computational cost in seconds for different configurations of APF and HMS.

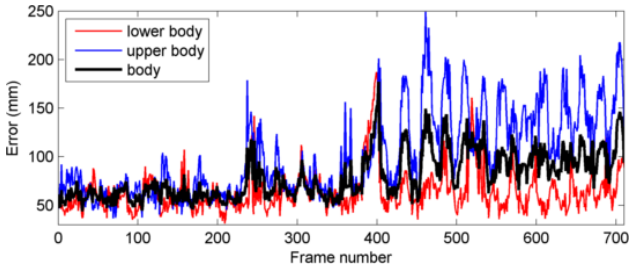

Figure 14: Average error per frame for HE-II S2 dataset with HMS(1-5) for lower body (red line) and upper body (blue line) and full body (black line).

In the Figures 15, 16 and 17 we display tracking results of HMS(1-5) with threshold 1 for the datasets used in the study.

5. Conclusions

This paper presents a human pose tracking methodology relying on two novel techniques. Firstly, a hierarchical method based on dimensionality reduction for human pose tracking is proposed. The hierarchical dimensionality reduction method, HTLE, based on TLE dimensionality reduction method, has been designed for human pose tracking as it takes into account the hierarchical representation of the human body. This allows the decoupling from the structure of the training dataset and the exploration of unseen poses.

Secondly, we introduce a method, HMS, which deterministically searches

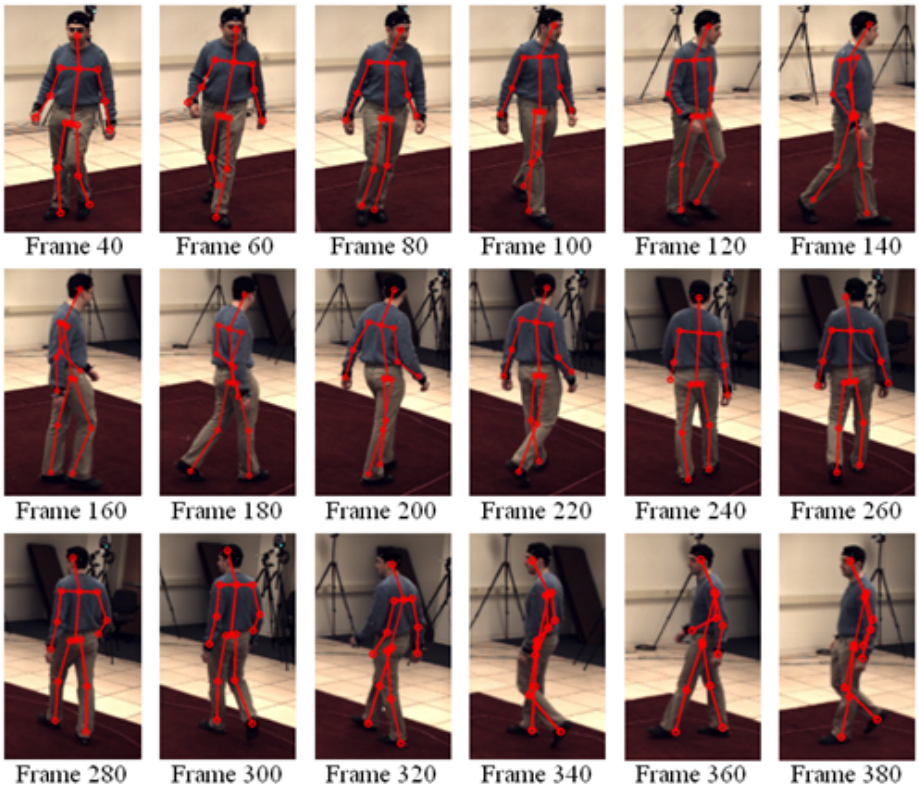

Figure 15: Results for HE-II S2 walking dataset with HMS(1-5).

through the hierarchy of low dimensional manifolds and is driven by an observation function. HMS allows searching in a constrained space for every level of the hierarchy, so it requires a low number of evaluations of the observation functions and therefore low computational resources. In addition, searching through the hierarchy is able to estimate unseen poses.

Results were presented using publicly available benchmarks, such as the multi-camera HumanEva . Comparisons with state-of-the-art methods demonstrate the accuracy and efficiency of our approach.

Amin, S., Andriluka, M., Rohrbach, M., Schiele, B., 2013. Multi-view Pictorial Structures for 3D Human Pose Estimation. Procedings of the British Machine Vision Conference 2013, 45.1–45.11.
URL http://www.bmva.org/bmvc/2013/Papers/paper0045/index.html

Arulampalam, M., Maskell, S., Gordon, N., Clapp, T., Sci, D., Organ, T.,

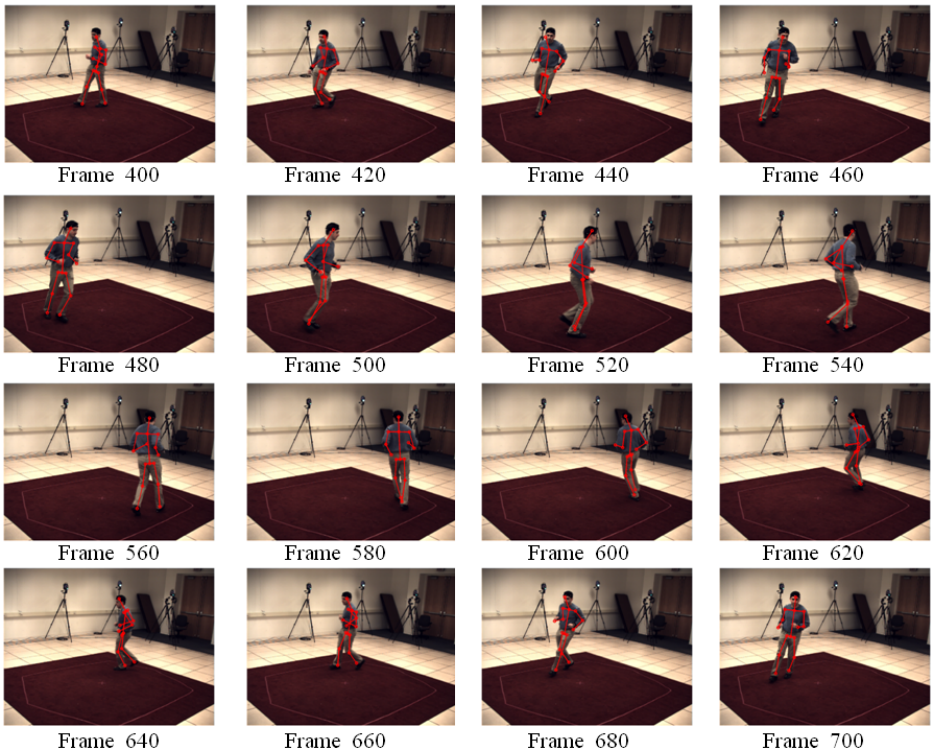

Figure 16: Results for HE-II S2 jogging dataset with HMS(1-5).

Adelaide, S., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions On Signal Processing 50 (2), 174–188.

Balan, A., Sigal, L., Black, M., 2005. A quantitative evaluation of video-based 3D person tracking. In: International Workshop on Performance Evaluation of Tracking and Surveillance. IEEE, pp. 349–356.

Bandouch, J., Engstler, F., Beetz, M., 2008. Evaluation of hierarchical sampling strategies in 3d human pose estimation. In: Proceedings of the 19th British Machine Vision Conference (BMVC).

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in neural information processing systems 14, 585–591.

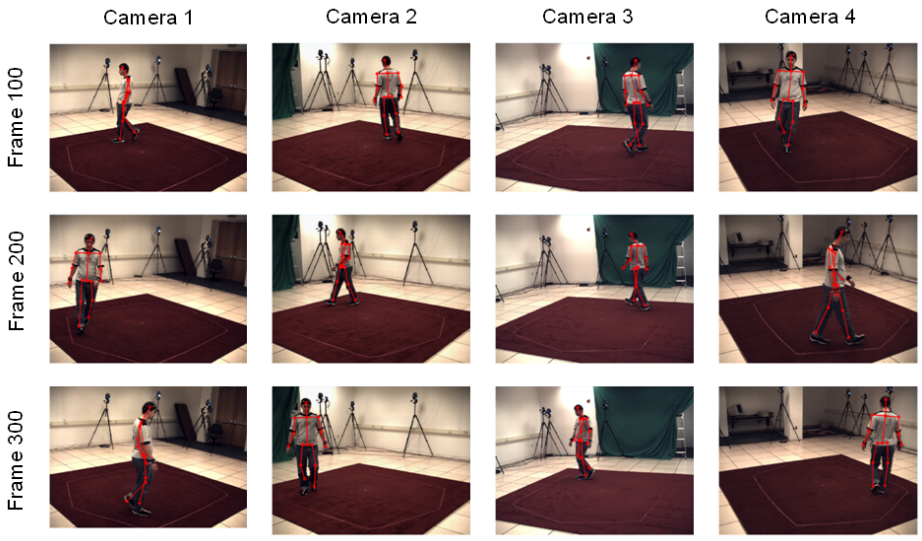

Figure 17: Results for HE-II S4 dataset with HMS(1-5) for four cameras.

Belkin, M., Niyogi, P., 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15 (6), 1373–1396.

BrownUniversity, 2007. In: HumanEva. [Online; accessed 27-December-2012].
URL http://vision.cs.brown.edu/humaneva/submit_results.html

Cheung, K., Baker, S., Kanade, T., 2005. Shape-from-silhouette across time part i: Theory and algorithms. International Journal of Computer Vision 62 (3), 221–247.

Cheung, K., Others, 2003. Visual hull construction, alignment and refinement for human kinematic modeling, motion tracking and rendering (October).
URL http://dl.acm.org/citation.cfm?id=997585

Darby, J., Li, B., Costen, N., Fleet, D., Lawrence, N., 2009. Backing off: Hierarchical decomposition of activity for 3d novel pose recovery. In: British Machine Vision Conference. Vol. 186. pp. 187–191.

Deutscher, J., Reid, I., Feb. 2005. Articulated Body Motion Capture by Stochastic Search. International Journal of Computer Vision 61 (2), 185–205.

Fitzgibbon, A., Cross, G., Zisserman, A., 1998. Automatic 3D model construction for turn-table sequences. 3D Structure from Multiple Images of Large-Scale Environments, 155–170.

Gall, J., Rosenhahn, B., Brox, T., Seidel, H.-P., Nov. 2008. Optimization and Filtering for Human Motion Capture. International Journal of Computer Vision 87 (1-2), 75–92.
URL http://link.springer.com/10.1007/s11263-008-0173-1

Han, L., Wu, X., Liang, W., Hou, G., Jia, Y., May 2010. Discriminative human action recognition in the learned hierarchical manifold space. Image and Vision Computing 28 (5), 836–849.
URL http://linkinghub.elsevier.com/retrieve/pii/S0262885609001723

Hou, S., Galata, A., Caillette, F., Thacker, N., Bromiley, P., 2007. Real-time body tracking using a gaussian process latent variable model, 1–8.

Howe, N. R., Leventon, M. E., Freeman, W. T., 1999. Bayesian reconstruction of 3d human motion from single-camera video, 820–826.

Jenkins, O. C. O., Matarić, M. M. J., 2004. A spatio-temporal extension to isomap nonlinear dimension reduction. In: Proceedings of the twenty-first international conference on Machine learning. ACM New York, NY, USA, ACM, p. 56.

Lawrence, N., 2004. Gaussian process latent variable models for visualization of high dimensional data. Advances in neural information processing systems 16, 329–336.

Lawrence, N. D., Moore, A. J., 2007. Hierarchical Gaussian process latent variable models. Proceedings of the 24th international conference on Machine learning - ICML '07, 481–488.
URL http://portal.acm.org/citation.cfm?doid=1273496.1273557

Lewandowski, M., Makris, D., Nebel, J.-C., Sep. 2010a. Automatic configuration of spectral dimensionality reduction methods. Pattern Recognition Letters 31 (12), 1720–1727.

Lewandowski, M., Makris, D., Velastin, S., Nebel, J.-C., June 2014. Structural laplacian eigenmaps for modeling sets of multivariate sequences. Cybernetics, IEEE Transactions on 44 (6), 936–949.

Lewandowski, M., Martinez-del Rincon, J., Makris, D., Nebel, J.-C., Aug. 2010b. Temporal Extension of Laplacian Eigenmaps for Unsupervised Dimensionality Reduction of Time Series. 2010 20th International Conference on Pattern Recognition, 161–164.

Moutzouris, A., del Rincón, J. M., Nebel, J., Makris, D., 2012. Human Pose Tracking by Hierarchical Manifold Searching. International Conference on Pattern Recognition (ICPR), 866–869.

Moutzouris, A., Martinez-del Rincon, J., Lewandowski, M., Nebel, J.-C., Makris, D., Sep. 2011. Human pose tracking in low dimensional space enhanced by limb correction. 2011 18th IEEE International Conference on Image Processing, 2301–2304.

Raskin, L., Rudzsky, M., 2009. Using Hierarchical Models for 3D Human Body-Part Tracking. Proceedings of the British Machine, 11–20.

Raskin, L., Rudzsky, M., Rivlin, E., Apr. 2011. Dimensionality reduction using a Gaussian Process Annealed Particle Filter for tracking and classification of articulated body motions. Computer Vision and Image Understanding 115 (4), 503–519.

Roweis, S. T., Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Sigal, L., Balan, A., Black, M., Aug. 2010. HumanEva : Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. International Journal of Computer Vision 87, 4–27.

Sigal, Leonid, 2004. Image & MOCAP Synchronized Dataset(v.1.0). In: Computer Vision and Pattern Recognition. [Online; accessed 27-December-2012].
URL `http://www.cs.brown.edu/ ls/Software/index.html`

Stauffer, C., Grimson, W., 1999. Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. Vol. 2. IEEE.

Tenenbaum, J. B., de Silva, V., Langford, J. C., Dec. 2000. A global geometric framework for nonlinear dimensionality reduction. Science (New York, N.Y.) 290 (5500), 2319–23.

Tian, Y., Zitnick, C. L., Narasimhan, S. G., 2012. Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation.

Urtasun, R., Darrell, T., 2008. Sparse probabilistic regression for activity-independent human pose inference. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Urtasun, R., Fleet, D., Fua, P., 2006. 3D people tracking with Gaussian process dynamical models. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 1. IEEE, pp. 238–245.

Urtasun, R., Fleet, D., Lawrence, N., 2007. Modeling human locomotion with topologically constrained latent variable models. Human Motion–Understanding, Modeling, Capture and Animation, 104–118.

Wang, J. M., Fleet, D. J., Hertzmann, A., 2006. Gaussian process dynamical models. Advances in neural information 18, 1441–1448.

Wang, Y., Tran, D., Liao, Z., Jun. 2011. Learning hierarchical poselets for human parsing. Cvpr 2011, 1705–1712.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5995519

Yang, H.-D., Lee, S.-W., 2006. Reconstructing 3d human body pose from stereo image sequences using hierarchical human body model learning. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 3. IEEE, pp. 1004–1007.

Zhang, Z.-y., Zha, H.-y., 2004. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. Journal of Shanghai University (English Edition) 8 (4), 406–424.